countries were the following: (1) keep the same date, event, and consequences whenever possible (i.e., Tuesday May 4, a 3 alarm fire, destruction of two hotels and one restaurant); and (2) substitute the place where the event is located [i.e., a city (Cleveland, Ohio)] with a place familiar to the subjects living in the target countries. **RESULTS:** The event (fire) could be kept in all countries. The date had to be changed in The Netherlands because it corresponds to a commemoration (Remembrance of the Dead) and would have introduced a bias if kept. The verbatim "a 3 alarm fire" was impossible to translate literally since no equivalent fire-classification system is used in most target countries (except in Canada). It was decided to use synonyms of "big" to qualify "fire." Syntax was also an issue especially in Korea, Japan, Romance and Germanic languages where the order of some segments had to be inverted. **CONCLUSIONS:** Although simple in its structure, the RBANS story memory-test proved to be challenging to translate into 24 languages and required a rigorous methodology to preserve the intent of the original.

PRM110

STANDARDIZATION OF MENTAL HEALTH ASSESSMENT – USING ITEM RESPONSE THEORY (IRT) TO CROSS-CALIBRATE TWO SELF-REPORTED MENTAL HEALTH TOOLS: THE PATIENT HEALTH QUESTIONNAIRE (PHQ-9) AND THE SF-36V2 MENTAL HEALTH (MH) SCALE

Bjorner JB, White MK, Yarlas AS

Optum, Lincoln, RI, USA

OBJECTIVES: Mental health can be measured by numerous instruments, but scores are usually not directly comparable. The heterogeneity of scale specific metrics seriously impairs comparability across study results and the communication among researchers and clinicians. We aimed to develop and evaluate methods for cross-calibration of two popular mental health tools: the PHQ-9 and the SF-36v2 MH scale. **METHODS:** We analyzed data from the United States and the UK including a general population sample (US: 216, UK: 355) and a sample with suspected depression (US: 169, UK: 153). Multigroup confirmatory bifactor models tested whether the two instruments measured the same construct. Differential item function (DIF) between general population and depression samples was tested using logistic regression DIF tests. We estimated IRT item parameters using a multigroup generalized partial credit model and developed cross-calibration algorithms using the summed score cross-calibration approach. The measurement properties of the instruments were evaluated by test information functions. **RESULTS:** In the bifactor model, all items loaded strongly on a common factor, supporting that the two scales measure the same general mental health construct. We found no indication of DIF, supporting that the same item parameters apply to the general population and the depression samples. The cross-calibration algorithm revealed a fairly linear relation between PHQ-9 score and MH score in the PHQ-9 score range of 10-20 (moderate to severe depression), but a non-linear relation at more extreme scores. The PHQ-9 provided most information for persons with scores in the interval from the general population average down to two standard deviations below average, but the MH scale provided more information at the lower and upper extremes. **CONCLUSIONS:** We successfully developed a procedure for cross-calibrating the PHQ-9 and MH scales. These results can be used to compare scores between the two instruments.

PRM111

INTERNAL VALIDATION OF MAPPING ANALYSES FOR HEALTH TECHNOLOGY ASSESSMENT

Trueman D, Treharne C

Abacus International, Bicester, UK

OBJECTIVES: Mapping between health status measures is common practice within health economic evaluations. The objective of this analysis was to evaluate the suitability of hold-out validation, whereby models are fitted to a subset of data and then tested in the remaining observations, compared to other methods of internal validation utilising full sample approaches in small to medium sized samples. **METHODS:** Four models predicting EQ-5D from the SF-12 were estimated using the Medical Expenditure Panel Survey. Models were estimated using three hypothetical sample sizes of 500, 1,000, and 4,000 observations. For each model and sample size, two hold-out validation specifications were compared against alternative estimators of error: the naïve resubstitution error; repeated 10-fold cross validation; the optimism-corrected bootstrap; the 0.632 bootstrap. The results from these estimators were compared against asymptotic estimates of the true error indices in the remaining observations (n=15,675). Estimators were evaluated by assessment of bias and variance. The exercise was repeated 500 times. **RESULTS:** Hold-out methods were subject to the largest variance across all estimators and sample sizes. Variance was lower and similar in the full sample estimators (bootstrap and cross-validation methods). The extent of bias in any sample size was associated with the degree to which the algorithms were adaptive to the training sample data. For the two mapping algorithms which were not adaptive to the training sample data, bias was low for all estimators. In the two algorithms which were more adaptive to the training sample data, the naïve resubstitution error was associated with a downward bias, hold-out methods exhibited an upward bias, and all full sample methods exhibited a low degree of bias. **CONCLUSIONS:** Hold-out validation exhibited the highest variance of all methods in all scenarios. Full-sample designs offer lower variance and are preferable to continued use of hold-out validation with small to medium sized datasets.

PRM112

A SYSTEMATIC REVIEW OF METHODOLOGICAL FRAMEWORKS FOR EVALUATION OF ETHICAL CONSIDERATIONS IN HEALTH TECHNOLOGY ASSESSMENT

Assasi N, Schwartz L, Tarride JE, Campbell K, Goeree R

McMaster University, Hamilton, ON, Canada

OBJECTIVES: While advances have been made in development of ethical frameworks for health technology assessment (HTA), there is no clear agreement on the most useful and practical approach to address ethical aspects in HTA. Moreover, uncertainty remains about appropriate scope and level of details of ethical frameworks for HTA. This study seeks to systematically review the literature to identify existing ethics frameworks for HTA in order to provide an overview of their methodological features and to gain a better understanding of the areas of commonality and divergence between different frameworks. **METHODS:** We conducted a systematic search of literature, without limits of time and language, to identify the guidance documents or practical frameworks published up to October 1st 2013. **RESULTS:** The review identified 22 frameworks, varying in their philosophical approach, structure, and comprehensiveness. They were designed for different purposes throughout the HTA process, ranging from helping HTA producers in identification, appraisal and analysis of ethical data to supporting decision-makers in making better informed value-sensitive decisions. They frequently promoted analytical methods that combined normative reflection with descriptive approaches to the analysis of values of stakeholders and other societal or technical actors. **CONCLUSIONS:** The choice of a method for collection and analysis of ethical data seems to depend on the context in which technology is being assessed, the purpose of analysis, and availability of required resources.

RESEARCH ON METHODS – Statistical Methods

PRM113

COMPARING PROPENSITY SCORE, PROPENSITY SCORE WITH COVARIATES AND GENETIC ALGORITHM METHODS FOR COVARIATE MATCHING IN OBSERVATIONAL STUDIES

Claeys C, Bakken DG, Wasserman D, Spilman J

KJT Group, Inc., Honeoye Falls, NY, USA

OBJECTIVES: As the population ages an increasing number of individuals are providing informal (unpaid) care for an aging relative. We compare three different methods of covariate matching to determine the effect of caregiving on the mental health states of informal caregivers. Covariate matching methods pair observations from different treatment groups by matching the members of each pair on a set or vector of covariates that would be randomly distributed across the groups in a randomized trial. **METHODS:** Multiple waves of an online survey conducted among a representative sample of U.S. adults yielded 740 informal caregivers and 2260 non-caregivers. We applied three different methods for covariate matching to determine the "average effect of treatment on the treated" (ATT) of caregiving on mental health states (MH): 1. Propensity score within calipers; 2. Propensity score and covariates within calipers; and 3. Genetic algorithm matching. **RESULTS:** All three methods provide adequate balance on the covariates used for matching. Methods 2 and 3 produce the best covariate balance, with absolute mean covariate differences less 0.0008 on all covariates and less than 0.00001 on the core set of covariates. Because methods that censor observations (i.e. matching within calipers) may artificially improve covariate balance, we take the ATT estimate from genetic matching to be the least biased estimate of the true effect. Using a standard 5-point self-report measure of mental health, caregivers, on average, report a mental health state that is 5.4% worse than non-caregivers (roughly one-fourth "less healthy" within any given scale range (e.g. 2-3, 3-4). **CONCLUSIONS:** As all three methods provide adequate matching, our consideration turns to bias reduction and the fact that the genetic matching does not require that we estimate the propensity score prior to matching. We consider the drivers or caregiver MH and implications for health care policy.

PRM114

ARE INDUSTRY FUNDED NETWORK META-ANALYSES LOWER QUALITY?

Chambers JD[1], Gunjal SS[2], Winn A[3], Kennedy IR[4], Hoey MG[5], Pyo J[1]

[1]*Tufts Medical Center, Boston, MA, USA,* [2]*University of Houston, Houston, TX, USA,* [3]*The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA,* [4]*Daisy Hill Hospital, Newry, UK,* [5]*Downe Hospital, UK*

OBJECTIVES: To compare the quality and transparency of industry supported network meta-analyses with those with non-profit support or no support. **METHODS:** We systematically searched OVID-Medline for network meta-analyses including at least one pharmaceutical. We reviewed each network meta-analysis and evaluated key general study characteristics, methodology, and transparency using a checklist of objective criteria derived from the ISPOR Taskforce's recommendations for study conduct and reporting. We reported source of study funding when available. When source of funding was unclear or not reported we contacted the corresponding author. We compared the quality and transparency of industry supported network meta-analyses with those with non-profit support or no support. **RESULTS:** Two hundred and fourteen studies met our inclusion criteria and were included in our dataset. Source of funding was identified for 211 studies (98.6%). Industry supported studies tended to be published in lower quality medical journals ($p<0.01$), and typically included fewer studies ($p<0.05$) and a smaller total number of patients ($p<0.05$). In terms of study transparency, industry supported studies less often reported the search terms ($p<0.01$) and, for analyses conducted using a Bayesian framework, presented the model code ($p<0.01$). Regarding study methodology, industry supported network meta-analyses less often reported a quality assessment of clinical studies included in the network meta-analysis ($p<0.01$), and less often compared the findings of traditional meta-analysis and network meta-analysis ($p<0.01$). With respect to presentation of findings, industry supported studies less often reported the full matrix of head-to-head comparisons ($p<0.01$), or provided a ranking of treatments ($p<0.01$). **CONCLUSIONS:** We found that studies with non-profit support or no support funded tended to be more transparent and rigorous than industry supported studies. Study findings emphasize that users of network meta-analyses should take great care to account for study quality when interpreting the findings of network meta-analyses.

PRM115

AUTOMATIC DEVELOPMENT OF CLINICAL PREDICTION MODELS WITH GENETIC PROGRAMMING: A CASE STUDY IN CARDIOVASCULAR DISEASE

Bannister CA, Currie CJ, Preece A, Spasic I

Cardiff University, Cardiff, UK

OBJECTIVES: Genetic programming is an Evolutionary Computing technique, inspired by biological evolution, capable of discovering complex non-linear patterns in large datasets. Despite the potential advantages of genetic programming over standard statistical methods, its applications to survival analysis are at best rare, primarily because of the difficulty in handling censored data. The aim of this study was to demonstrate the utility of genetic programming for the automatic development of clinical prediction models using cardiovascular disease as a case study. **METHODS:** We compared genetic programming and the commonly used Cox regression technique in the development of a cardiovascular risk score using data from the SMART study, a prospective cohort study designed to identify predictors of future cardiovascular events in patients with symptomatic cardiovascular disease. The primary outcome was any cardiovascular event, comprising cardiovascular death and non-fatal stroke and myocardial infarction. The predictive ability of the model was assessed in terms of discrimination and calibration. **RESULTS:** 3,873 patients were enrolled in the study 1996–2006, aged 19–82 years and with 460 cardiovascular events. The discrimination of both models was comparable; the C-index of the genetic programming model being smaller (0.65; 95% CI: 0.63–0.66) but not significantly different from that of the Cox regression model (0.71; 0.67–0.75). The calibration of both models was also comparable, indicating similar disagreement between observed and predicted risks. **CONCLUSIONS:** Using empirical data, we demonstrated that a prediction model developed by the novel technique of genetic programming has a comparable predictive ability to that of Cox regression. The genetic programming model was more complicated but was developed in an automated fashion and did not require the expertise needed for survival analysis. Genetic programming seems a promising technique for the automated development of clinical prediction models for diagnostic and prognostic purposes.

PRM116

IMPROVED BOOTSTRAP POINT AND CONFIDENCE INTERVAL ESTIMATION OF THE INCREMENTAL COST-EFFECTIVENESS RATIO (ICER)

Skrepnek GH[1], Sahai A[2]

[1]*The University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA,* [2]*The University of The West Indies, St. Augustine, Trinidad and Tobago*

OBJECTIVES: To develop and test a novel approach to estimate the ICER via a new bootstrap approach based upon the sample coefficient of variance and optimized via computational intelligence. **METHODS:** A novel bootstrap ICER estimation approach was developed that incorporated the sample coefficient of variance to better capture information within cost-effectiveness data. In this derivation, an optimal design value parameter was also obtained via computational intelligence. Across illustrative cost and outcome correlation structures and sample sizes, a simulation study of 1111 replications with 999 bootstrap resamples each was conducted utilizing MatLab R2012b. Comparative results of point estimates versus the existing bootstrap method were presented as relative efficiencies, with 95% confidence intervals (CI) presented as coverage probability, coverage error, length, left and right bias, and relative bias. **RESULTS:** The proposed ICER yielded less statistical estimation error than the typical bootstrap approach across all cases, with the relative efficiency of point estimates ranging from +106.03% to +113.35%. An equal or improved coverage error for the CI was also consistently achieved, deviating from the population value by zero (i.e., perfect coverage) to 0.0200 versus from 0.0060 to 0.0210. Subsequently, an improved shortening of the CI length was noted. The relative bias suggested slightly more left bias and less right bias across both positive and negative cost and outcome correlation structures, reaching a maximum of 0.5238 for the proposed ICER versus 0.2222 for the usual bootstrap. **CONCLUSIONS:** This novel method to estimate the ICER via the sample coefficient of variation found improvements in the relative efficiency of point estimators and in the coverage error and length of the 95% CI across all simulated cases. Irrespective of cost and outcome correlation structure, the relative bias of this ICER suggested a slight increase in potential left-sided bias and decrease in right-sided bias versus the usual bootstrap.

PRM117

USING MULTIPLE IMPUTATION FOR MISSING VALUES TO IDENTIFY CHRONIC KIDNEY DISEASE STAGES

Cai Y[1], Jiao X[2]

[1]*IMS Health, Plymouth Meeting, PA, USA,* [2]*IMS Health, Plymouthmeeting, PA, USA*

OBJECTIVES: Health care researchers often encounter missing values in many datasets. Ultimately, a patient record with missing fields can still carry valuable information. This extra information becomes more important to keep in oncology and other rare disease studies where sample size is typically limited. The purpose of this study is to demonstrate that researchers can benefit from using Multiple Imputation (MI) approach to tackle missing value problems. **METHODS:** The model data from IMS claims (Dx) and retail prescription (Rx) contained year 2011 patient level CKD stage indications, longitudinal drug therapies, days of supplies, titration rates, Demographic characteristics, payment type, and physician specialties, etc. We built multivariate logistic models to identify Chronic Kidney Disease (CDK) patient stages using prescription data in order to further evaluate the prevalence, economic burden and market opportunities. Under the general assumption of missing at random (MAR), we used MI with regression method to impute the missing monotone and categorical values before the modeling process. **RESULTS:** The pooled results from 5 MI imputed datasets were reported. Compared with the results from deterministic missing imputation approach, the MI showed larger standard error and wider 95% confidence interval. The wider CI reflected the additional data uncertainty from the missing values. CDK stage 4 (11.2%) had smallest proportion and it had lowest hit rate in the prediction model. MI approach showed more CDK stage 4 identifications than those from deterministic complete case analysis. **CONCLUSIONS:** This study demonstrated that MI is capable of reflecting the underline uncertainty associated with the data by introducing random errors into the imputation process. MI can generate unbiased results and good standard error estimation when using appropriately. On the other hand, with the advent of modern computational technology, the MI becomes computationally simple and easy to use.

PRM118

IDENTIFY CHF AND COPD PAITIENTS AT HIGH RISK OF HOSPITALIZATION: USING PREDICTIVE ANALYTICS FOR PATIENT OUTREACH

Wang QC, Higgins SL, Chawla R, Nigam S

Independence Blue Cross, Philadelphia, PA, USA

OBJECTIVES: To develop predictive models to identify CHF and COPD patients at high risk of hospitalization within the next 6 months to be used for case management outreach. **DATA SOURCES:** Data were extracted from several sources and included patient diagnoses, service utilization, lab data, and medication adherence from a large health insurance claims database, ZIP code level demographic data from the U.S. Census, patient level illness burden scores, medical episode groupers, Experian consumer and credit information, and call data between patients and customer service representatives. **STUDY POPULATION:** All commercial and Medicare members who were identified with congestive heart failure (CHF) and chronic obstructive pulmonary disease (COPD) and who were continuously enrolled for at least 6 months during the model development period and 6 months during the predictive period were included. **METHODS:** Using three years of historical data from 2010 to 2012 and admissions between January and June 2013 as the target variable, the data were randomly split in half as training and validation data. The training data were used to build the predictive model. The validation data were used to evaluate model performance. Several algorithms were utilized to build predictive models: logistic regression, neural networks, and decision trees. The models were evaluated based on the lift chart and/or area under the ROC curve. The selected models were used to score data and predict future admissions. **RESULTS:** The key factors predicting admissions in the next 6 months included length of time identified with CHF and COPD, medication adherence, prior admissions, recent specialist visits, having had a customer call that mentioned 'hospital bed/hospital stay', and being on oxygen (for COPD). **CONCLUSIONS:** Four models of predicting patients at highest risk of admission have been developed, which were used to generate a list of patients with high probability of admission for case management outreach.

PRM119

AN ANALYTICAL METHOD FOR ESTIMATING THE BOUNDARIES OF AN INCREMENTAL COST-EFFECTIVENESS RATIO

Kamae I[1], Yamabe K[2], Sugimoto T[1]

[1]*The University of Tokyo, Graduate School of Public Policy, Tokyo, Japan,* [2]*MSD.K.K. & HTA and Public Policy, Graduate School of Public Policy, The University of Tokyo, Tokyo, Japan*

OBJECTIVES: To develop an analytical method which quantifies the reasonable limits for any incremental cost-effectiveness ratio (ICER) defined by the slope of a line connecting two points on the cost-benefit plane. **METHODS:** Assume that the ICER of a target technology vs. its comparator is defined with two points at each of which a pair of cost and benefit is given on the C(cost)-E(benefit) plane. In order to find a cost-benefit function connecting the two points, an analytical method was developed by means of curve-fitting technique with exponential and quadratic modeling. The resultant cost-benefit function was further analytically expanded to the derivative, dC/dE, we call it "tangent limit". Example calculations of the tangent limits were conducted for each model. **RESULTS:** The analytical development resulted in the following equations of the cost-benefit function and the derivative for each modeling: $C = \mathrm{Exp}((E - p)/q)$ and $dC/dE = (1/q)\,\mathrm{Exp}((E - p)/q)$ for exponential model, whilst $C = (1/q)E^2 - p/q$ and $dC/dE = 2E/q$ for quadratic model, where p and q are parameters determined by costs and benefits of the target technology and its comparator. Applying the equations for two hypothetical points, (7.6 QALY, US$100,000) and (8.6 QALY, US$150,000), we found that the ICER of 50 bounds with the lower and the upper limits, respectively, 40.6 and 60.8 US$(x1000)/QALY for exponential model, and as well, 46.9 and 53.1 for quadratic model. Those estimates were not so much different as the limits of 43.9 and 65.8 for the ICER of 54.1, obtained by the regression analysis presented in the ISPOR New Orleans 2013. **CONCLUSIONS:** Our approach can offer a simple and science-based method to estimate boundaries for any ICER. It would be useful for negotiations and decisions in value-based pricing in which a range of ICER must be considered beyond a single threshold ratio.

PRM120

BIAS WHEN USING PROPENSITY SCORE METHODS TO ADJUST FOR COVARIATES THAT ARE NOT CONFOUNDERS

Chia VM, Page JH

Amgen, Inc, Thousand Oaks, CA, USA

OBJECTIVES: High-dimensional propensity score (PS) methods have been used in health care claims data to improve control of confounding by adjusting for a large number of covariates that may be proxies for unobserved factors. We have previously shown that PS models are biased for non-linear link functions when confounders were included. We conducted a simulation study to understand whether inclusion of covariates that are not confounders may also bias the association by estimating Monte Carlo mean bias, relative efficiency (RE) and coverage probability (CP) of log odds ratios when covariates only related to the exposure or only related to the outcome were included. **METHODS:** We conducted 1000 Monte Carlo simulations, and estimated effect of exposure using logistic regression models. The propensity score was included in the logistic model as a linear predictor or as a smoothed covariate using restricted cubic splines. Simulations were conducted for scenarios including 5, 15, and 25 covariates. **RESULTS:** Using the PS with 25 covariates related only to the binary exposure, Monte Carlo bias, standard error (SE), RE and CP were -0.002, 0.015, 1.34, and 0.94 when the PS was included as a smoothed covariate, and –0.002, 0.015, 1.31, and 0.94 when the PS was included as a linear covariate. The bias, SE, RE and CP for 25 covariates related to the binary outcome were 0.307, 0.096, 21.6, and 0 when the PS was included as a linear covariate. Bias tended to increase with more covariates. **CONCLUSIONS:** We observed minimal bias when using PS models where covariates were related only to the exposure, but substantial bias when the covariates were related to the outcome. PS models may not be appropriate for logistic models because these models do not adequately deal with errors in the outcome due to the covariate.